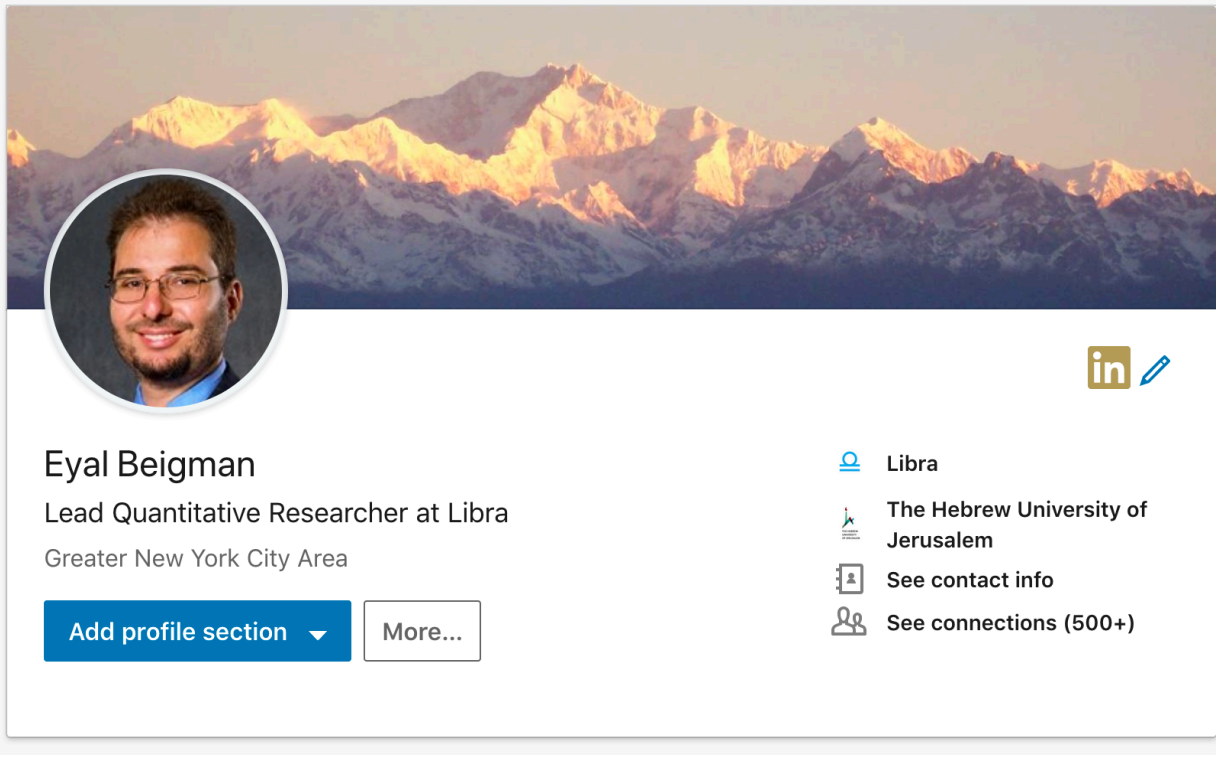
Data Science for QE

Eyal Beigman, PhD

About me


Eyal Beigman
Lead Quantitative Researcher at Libra
Greater New York City Area
Add profile section
More...
Libra
The Hebrew University of Jerusalem
See contact info
See connections (500+)

What is Data Science?

- *An **interdisciplinary** field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured*

  Wikipedia

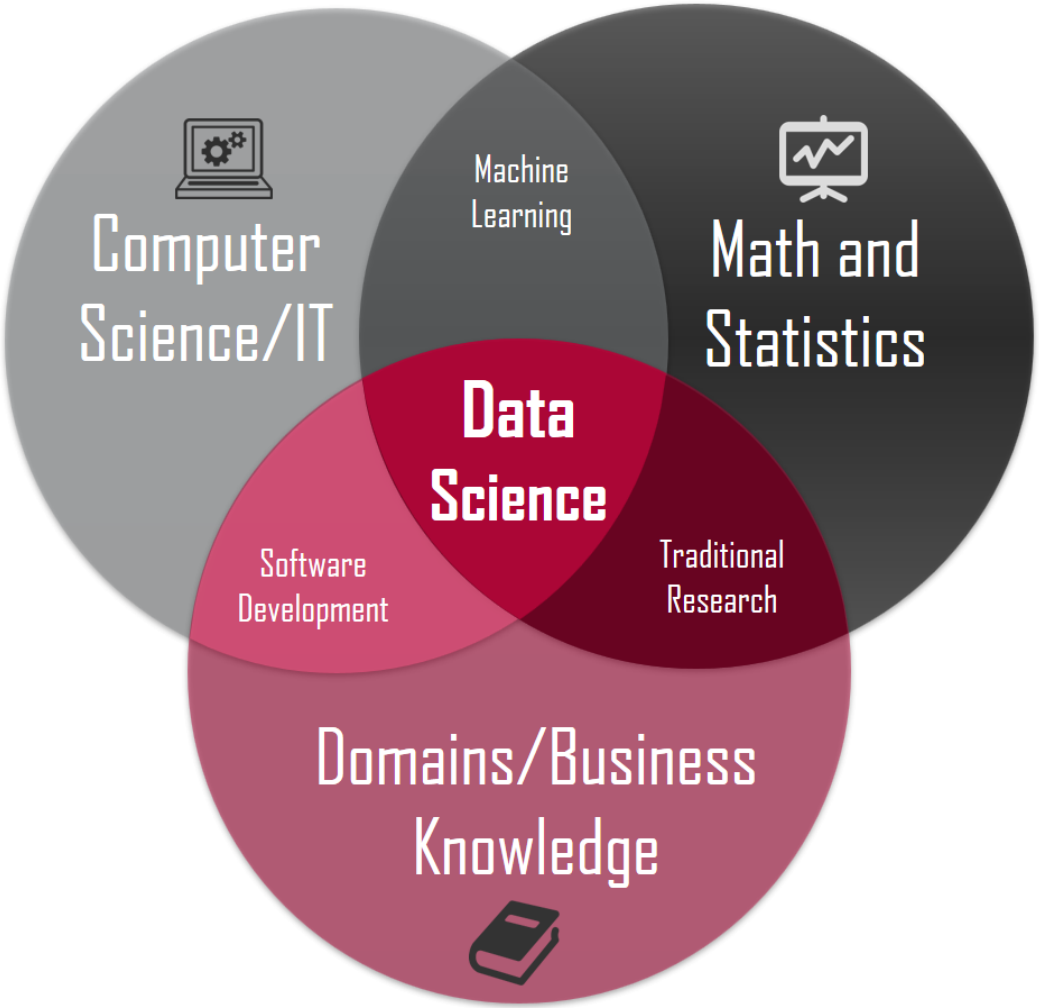
- *A **concept to unify** statistics, data analysis, machine learning and their related methods in order to understand and analyze actual phenomena with data*

  Chikio Hayashi

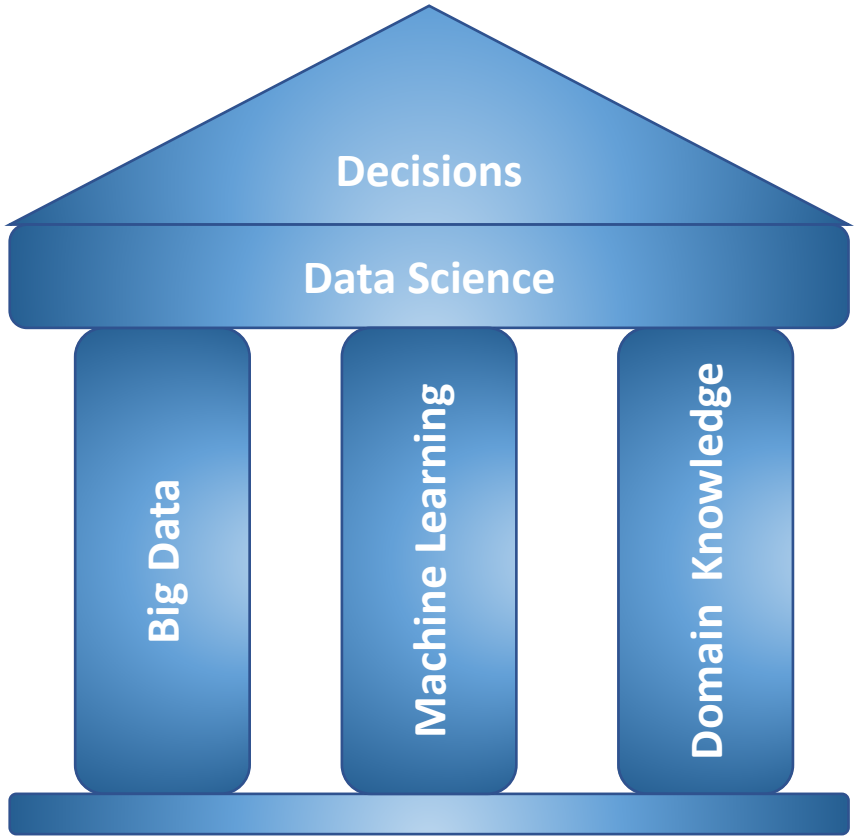
- *A sexed-up term for a statistician*

  Nat Silver

What is Data Science?



What is Data Science?



Brief History of Data Science

1962 - The Future of Data Analysis [Tukey]

1977 - Exploratory Data Analysis [Tukey]

1989 - *Knowledge Discovery in Databases (KDD)* [Workshop]

1994 – **Yahoo founded**

- Mining Data for Nuggets of Knowledge.

1996 - From Data Mining to Knowledge Discovery in Databases [Fayyad et al]

1997 - *Data Mining and Knowledge Discovery* [Journal]

*1998 – **Google founded***

2001 - Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics [Cleveland]

*2002 - **LinkedIn founded***

- Data Science Journal [Journal]

2003 – The Journal of Data Science [Journal]

- The Google File System [Google]

*2004 - **Facebook founded***

2005 - *Competing on Analytics* [Davenport, Cohen, Jacobson]

- Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century [NSA]

2008 – FIveThirtyEight founded

*2009 – "**the sexy job in the next ten years**" [Hal Varian, Chief economist Google]*

- The Revolution in Astronomy Education: Data Science for the Masses [Conference]

- Data Scientists group on LinkedIn

2010 - *A Taxonomy of Data Science*

- First Kaggle competition

2011 - ***Building Data Science Teams***

- First release of Hadoop

*2012 - **Data Scientist: The Sexiest Job of the 21st Century** [HBR]*

– FIveThirtyEight forecasts correctly election outcome in all fifty states

*2013 – **Data Science for Business***

Buzzwords

Data Science

Big Data

Hadoop

Business Intelligence

Data Lake

NoSQL

Datafication

MapReduce

Apache Sparc

SQL

Cloud computing

Machine Learning

Kdb+

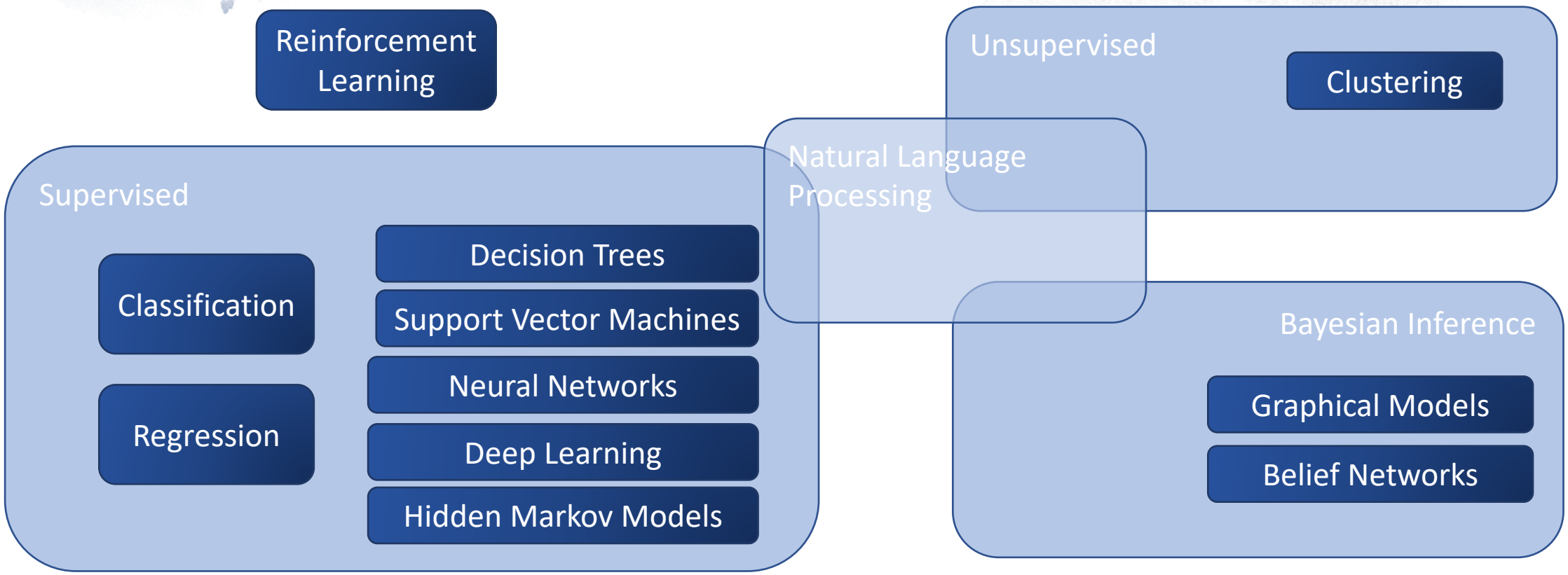
Data Mining

Postgres

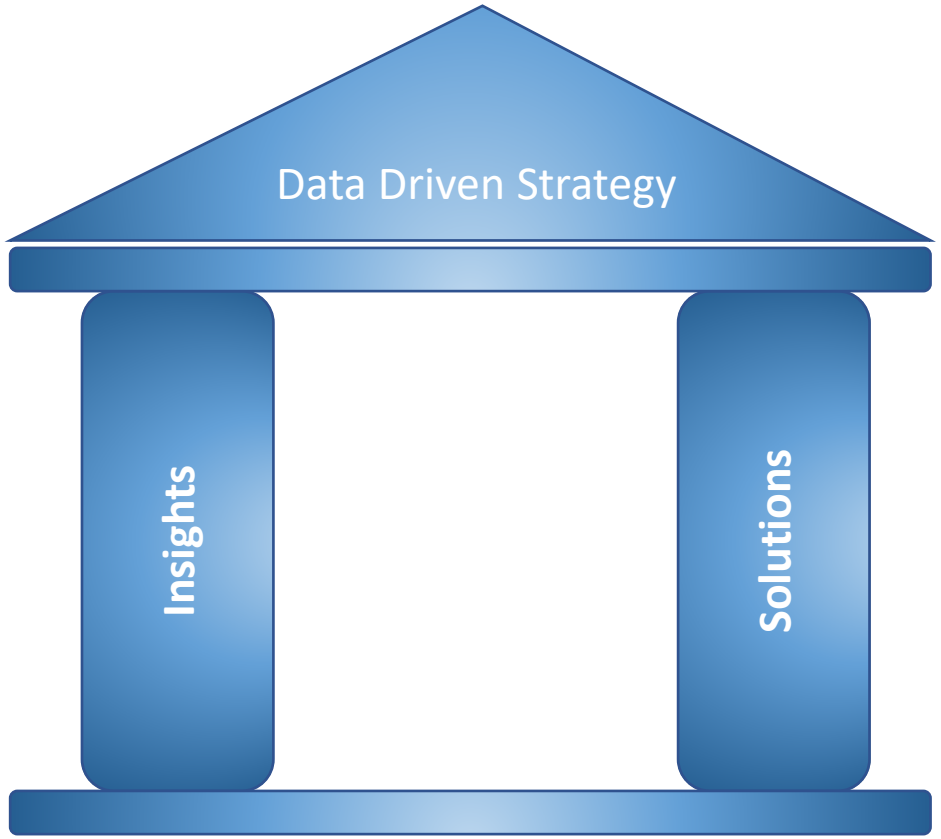
Deep Learning

Pattern Recognition

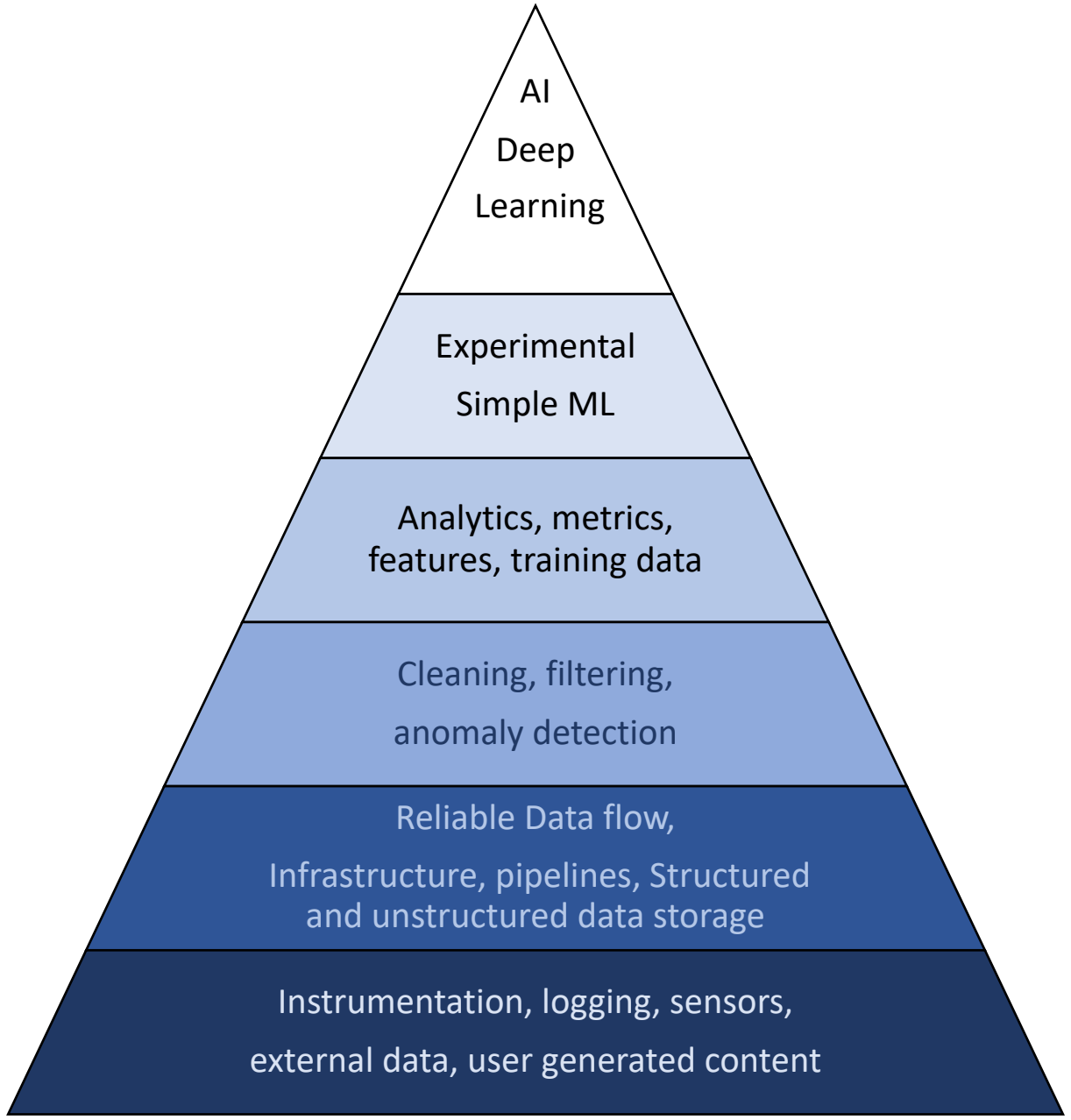
Machine Learning



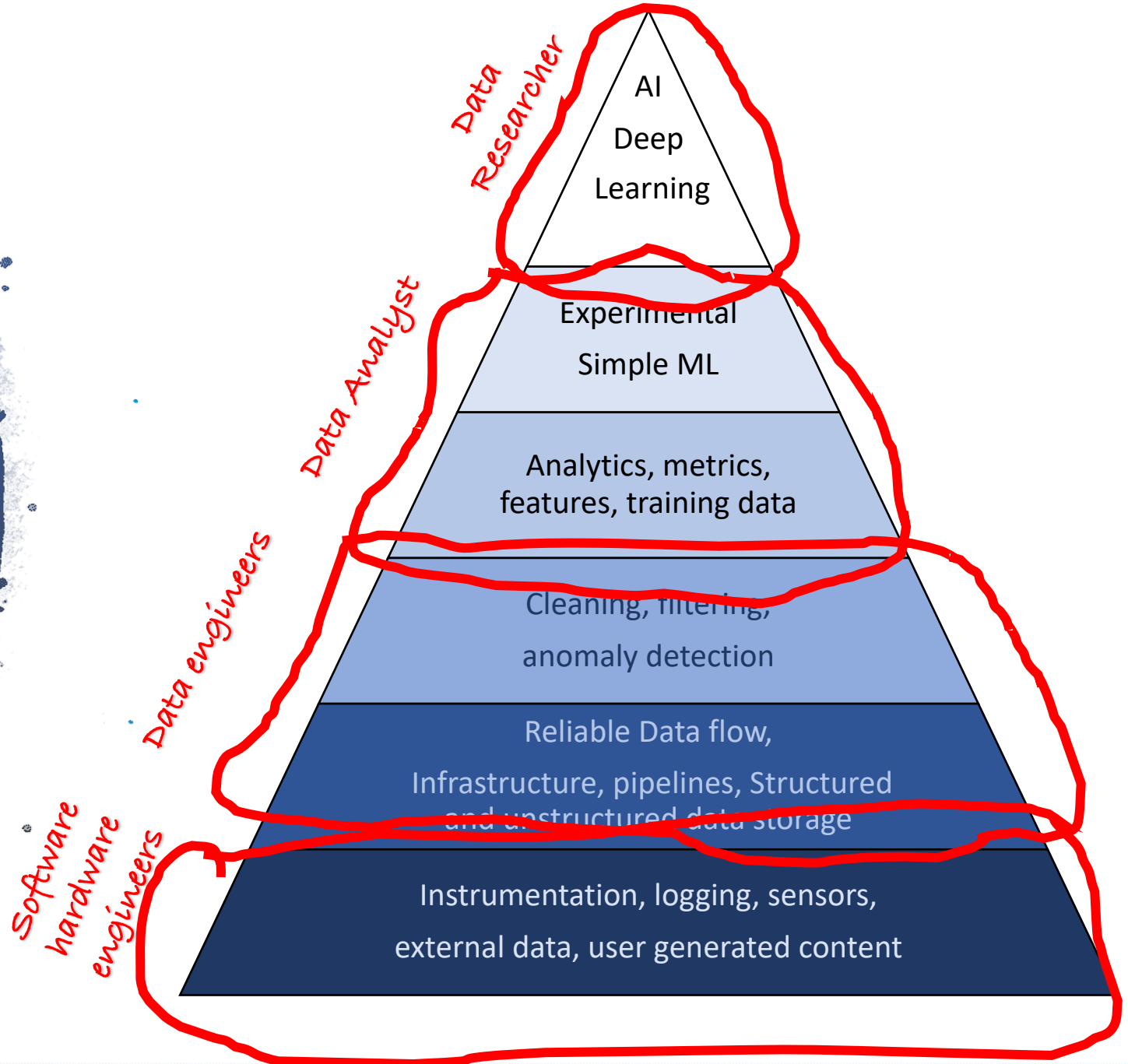
What Does Data Science do?



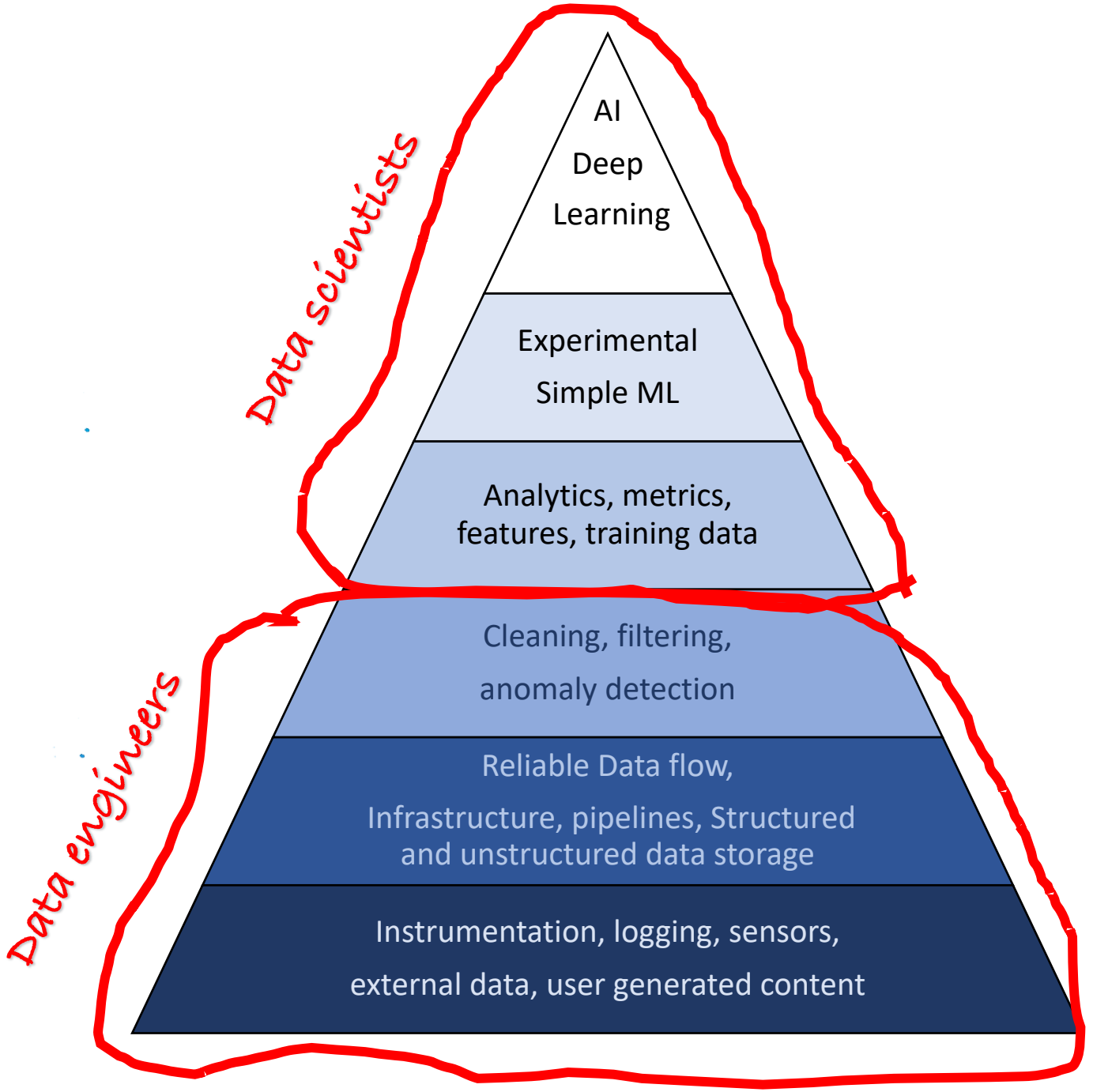
What Does a Data Scientist do?



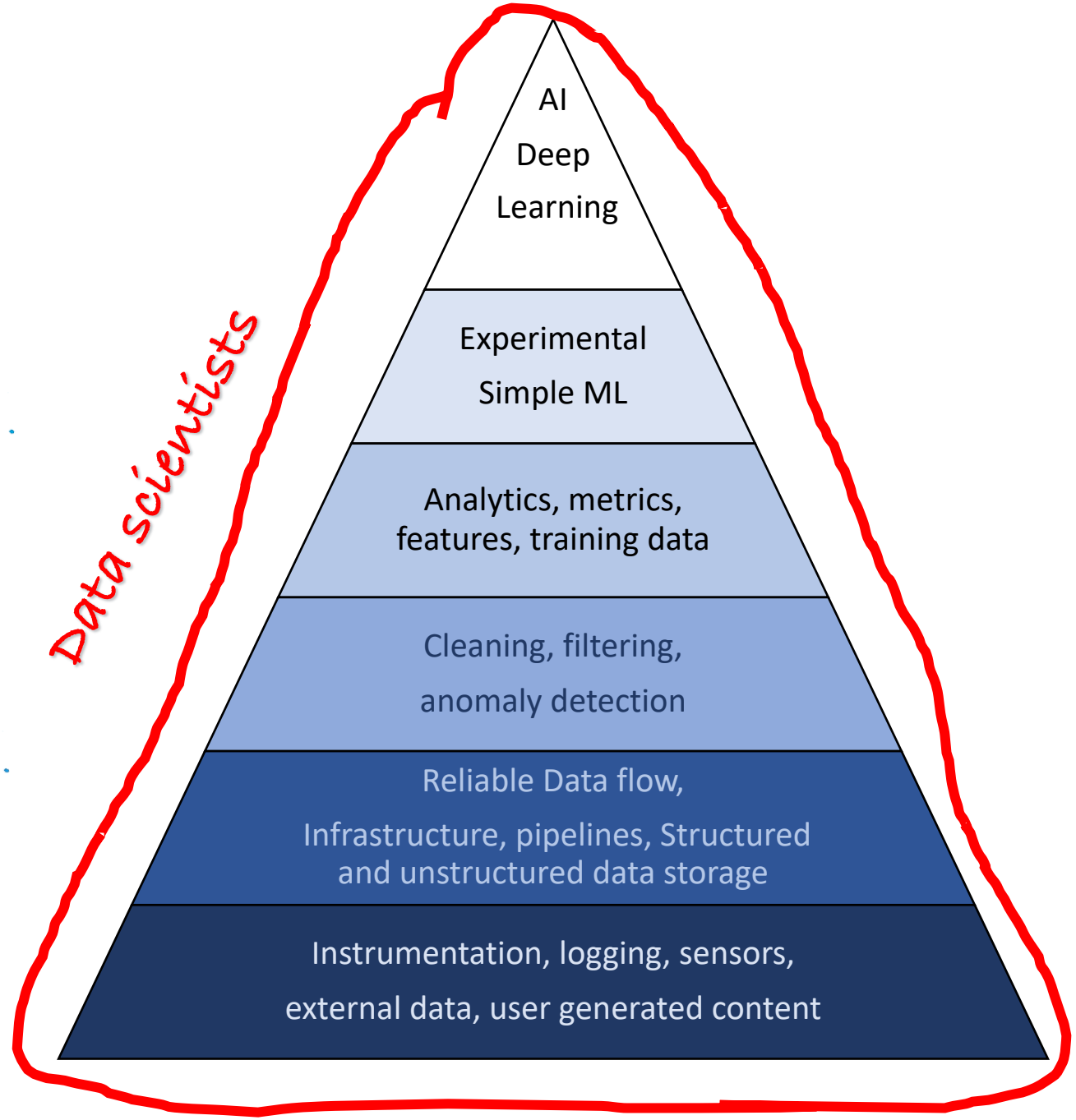
Large Organizations



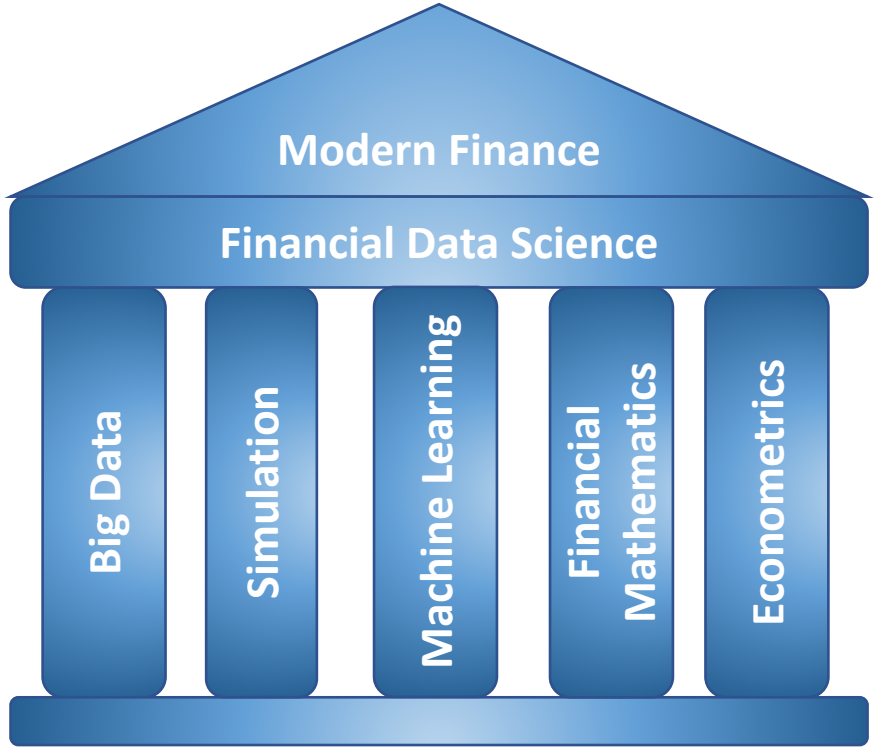
Medium Organizations



Small Organizations and Startups



Data Science in Finance



Special Characteristics

- Time Series
- Stochastic Process
- Real time

Data Scientists in the Finance Industry

- Economic/Fundamental Research (Analyst)
- Alpha Research (Research Quants)
- Algorithmic Trading (Algo Quants, Quant Developers)
- Risk Analysis (Strats)
- Asset Pricing (Desk Quants)
- Portfolio Management (Portfolio Manager, Quant Trader)
- Surveillance, Compliance and Fraud Detection (Analyst, Data Scientist)
- Data Support (Data Engineer, Data Architect)

Course Overview

Deflategate

- Week 1 : Introduction
- Week 2: Data Collecting

Exercise 1

Lending Club A

- Week 3 : Data Structure
- Week 4 : Data Cleaning

Exercise 2

Lending Club B

- Week 5 : Classification
- Week 6 : Regression
- Week 7 : Overfitting and avoidance

Exercise 3

- Week 8 : Model Evaluation

Course Overview

Midterm week 9

Week 9 : Unsupervised Learning

Blockchain and Bitcoin

Week 10 : Financial Data

Week 11: Bayesian Methods

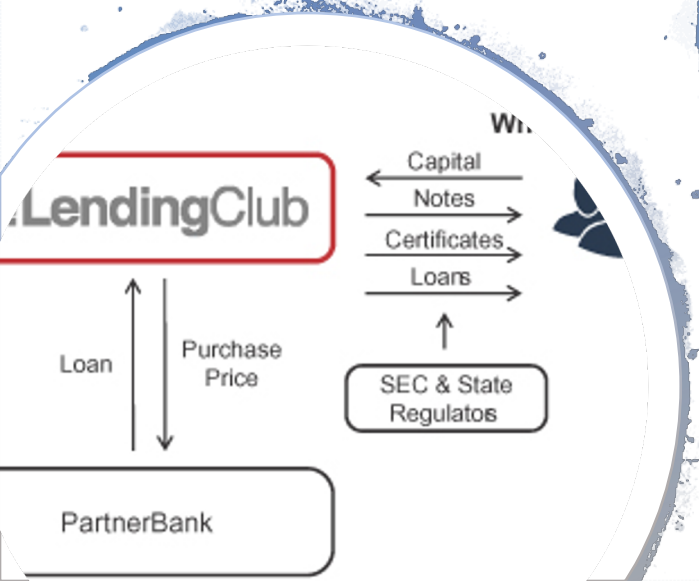
Week 12 : Text Analysis

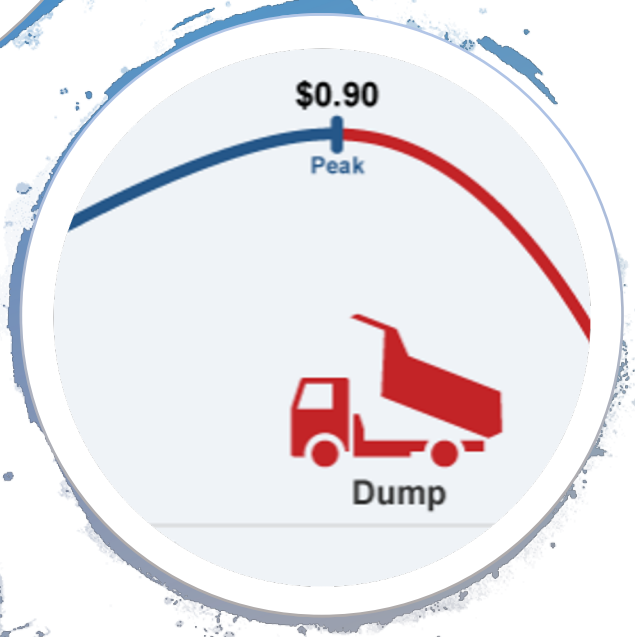
Exercise 4

Week 13 : Data Science and Business Strategy

Week 14 : Course Recap

Week 15 : Project Presentation





Cases Covered in the Course

- Deflategate
  - Scrapping data from on-line sources
    - RESTful API, XML, JSON
  - Organizing data in meaningful structures
- Lending Club
  - Cleaning and processing data
  - Classifiers
  - Data driven investment
- Pump and Dump
  - Processing exchange data
  - Patterns and outliers
  - Bayesian analysis
  - Data from text (news, chats and emails)

Case Studies

- Reading material for the course includes a HBS case pack
- The syllabus specifies the class in which each case will be featured
- You are expected to read the case ahead of class and come prepared
- At the start of the class there will be a class discussion on the case, you will be graded on your participation in this discussion
- The following link should be used to connect to HBS, register and purchase the pack
  - https://hbsp.harvard.edu/import/604441
- Cost should be $26.49

Exercises

- There will be four exercise throughout the course corresponding to the case studies
- Exercises will be published on Canvas and GitHub Classroom
- A Jupyter notebook template for each exercise will be published on GitHub Classroom
- The exercise will be submitted via GitHub Classroom
- All exercises should be done in Python using only the libraries in the template notebook

Final Project

- Working individually or in pairs, you will identify an interesting economic problem
- You will be asked to identify and compile a relevant dataset and make it available on GitHub
- You will perform an analysis of the problem using one or more of the methods covered in the course to gain insight into the problem, and offer a solution
- By the end of week 9 you need to submit an initial draft for the final project proposal for my vetting
- By end of week 11 you need to submit the final proposal
- Students will present their final projects at a “business level” to their peers

Course Bureaucracy

Assignment	Grading
Class Discussion	20%
Exercises (4x8)	32%
Mid Term Exam	23%
Final Project (18) and Presentation (12)	25%

Course Bureaucracy

Performance	Letter Grade	Range %
Excellent	A	93 - 100
	A-	90 - 92.9
Good	B+	87 - 89.9
Satisfactory	B	83 - 86.9
Below Average	B-	80 - 82.9
Poor	C+	77 - 79.9
	C	70 - 76.9
Failure	F	< 70

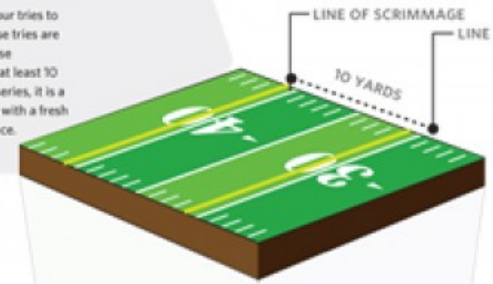
Deflategate – American Football

HOW TO PLAY AMERICAN FOOTBALL

The game of football is played with 22 players on the field, 11 players from each team. The **offense** is the team with the ball. The goal of the offense is to run plays to move the ball down the field to their end zone to score points. Plays are planned moves that the team performs together. The **defense** is the team without the ball. The goal of the defense is to stop the offense from scoring points, and to try to get the ball away from the offense. The team with the most **points** at the end of the game wins.

ADVANCING THE BALL

In one possession, the offense has four tries to move the ball forward 10 yards. These tries are called "downs".. Each time the offense successfully moves the ball forward at least 10 yards from where they started that series, it is a first down and players are rewarded with a fresh set of downs to continue their advance.



HOW TO PLAY DEFENSE

INTERCEPTION

Catching passes meant for players on the other team. Interceptions can be run to the end zone for touchdowns.

TACKLE

Bringing the player with the ball to the ground. Tackles are meant to stop the player with the ball and end the down.

HOW TO PLAY OFFENSE

RUN

When a player tries to carry the ball past the defense towards their endzone.

KICK

The offense may kick the ball towards the defense. This usually happens as a strategy during the 4th down if they are not close enough to score.

PASS

When a player throws the ball to a teammate. Only one forward pass can be made per down.

MISCELLANEOUS

FUMBLE

When a player drops a ball after having possession of it. Any player, offense or defense, can pick up the ball after it is fumbled. The team that picks it up first after the fumble then has possession.

SCORING

TOUCHDOWN

A touchdown occurs when the offense carries the ball into or catches the ball in their respective end zone. When this happens, the offensive team is awarded six points.

+6

2PT CONVERSION

After a touchdown, the offense gets one play to add either one or two extra points. Two extra points are scored by running or passing the ball into the end zone starting from the two-yard line.

+2

FIELD GOAL

The offense can score a field goal by kicking the ball through their goal posts and over the crossbar. If the goal is good, the offensive team is awarded three points.

+3

EXTRA POINT

After a touchdown, the offense gets one play to add either one or two extra points. One extra point is scored by kicking a field goal from the 2 yard line.

+1

SAFETY

The only time the defense can score. Safeties are scored when an offensive player: A steps of bounds in his own end zone, B is tackled, or C drops the ball.

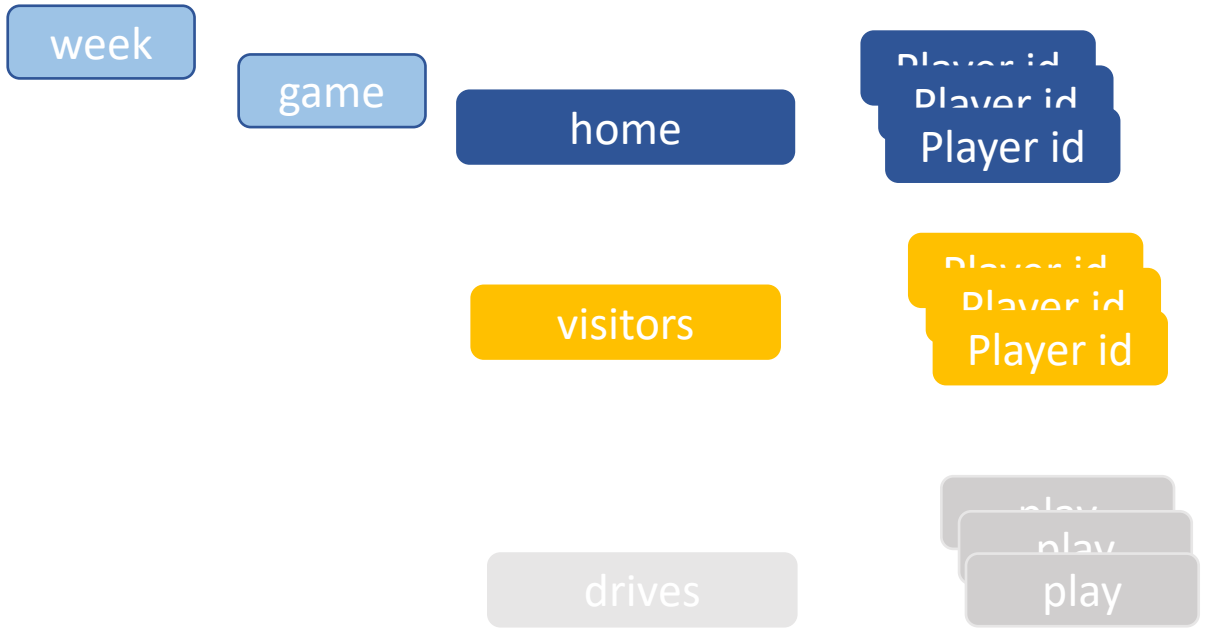
+2

PENTALTIES

Fouls are punished with penalties against the offending team. Most penalties result in moving the football towards the offending team's end zone.

Most penalties result in replaying the down. Some defensive penalties give the offense an automatic first down. Conversely, some offensive penalties result in loss of a down.

American Football Data



Weekly Games Reference

Query:

- http://www.nfl.com/ajax/scorestrip?season=**2018**&seasonType=**REG**&week=**16**


```
−<ss>
  −<gms gd="0" w="16" y="2018" t="R">
      <g eid="2018122200" gsis="57794" d="Sat" t="4:30" q="F" k="" h="TEN" hnn="titans" hs="25" v="WAS" vnn="redskins" vs="16" p="" rz="" ga="" gt="REG"/>
      <g eid="2018122201" gsis="57792" d="Sat" t="8:20" q="F" k="" h="LAC" hnn="chargers" hs="10" v="BAL" vnn="ravens" vs="22" p="" rz="" ga=""
      gt="REG"/>
      <g eid="2018122305" gsis="57796" d="Sun" t="1:00" q="F" k="" h="CLE" hnn="browns" hs="26" v="CIN" vnn="bengals" vs="18" p="" rz="" ga=""
      gt="REG"/>
      <g eid="2018122306" gsis="57797" d="Sun" t="1:00" q="F" k="" h="DAL" hnn="cowboys" hs="27" v="TB" vnn="buccaneers" vs="20" p="" rz="" ga=""
      gt="REG"/>
      <g eid="2018122307" gsis="57798" d="Sun" t="1:00" q="F" k="" h="DET" hnn="lions" hs="9" v="MIN" vnn="vikings" vs="27" p="" rz="" ga="" gt="REG"/>
      <g eid="2018122308" gsis="57799" d="Sun" t="1:00" q="F" k="" h="NE" hnn="patriots" hs="24" v="BUF" vnn="bills" vs="12" p="" rz="" ga="" gt="REG"/>
      <g eid="2018122309" gsis="57800" d="Sun" t="1:00" q="FO" k="" h="NYJ" hnn="jets" hs="38" v="GB" vnn="packers" vs="44" p="" rz="" ga="" gt="REG"/>
      <g eid="2018122310" gsis="57801" d="Sun" t="1:00" q="F" k="" h="PHI" hnn="eagles" hs="32" v="HOU" vnn="texans" vs="30" p="" rz="" ga="" gt="REG"/>
      <g eid="2018122304" gsis="57795" d="Sun" t="1:00" q="F" k="" h="CAR" hnn="panthers" hs="10" v="ATL" vnn="falcons" vs="24" p="" rz="" ga=""
      gt="REG"/>
      <g eid="2018122300" gsis="57791" d="Sun" t="1:00" q="F" k="" h="IND" hnn="colts" hs="28" v="NYG" vnn="giants" vs="27" p="" rz="" ga="" gt="REG"/>
      <g eid="2018122302" gsis="57793" d="Sun" t="1:00" q="F" k="" h="MIA" hnn="dolphins" hs="7" v="JAX" vnn="jaguars" vs="17" p="" rz="" ga=""
      gt="REG"/>
      <g eid="2018122311" gsis="57802" d="Sun" t="4:05" q="F" k="" h="ARI" hnn="cardinals" hs="9" v="LA" vnn="rams" vs="31" p="" rz="" ga="" gt="REG"/>
      <g eid="2018122312" gsis="57803" d="Sun" t="4:05" q="F" k="" h="SF" hnn="49ers" hs="9" v="CHI" vnn="bears" vs="14" p="" rz="" ga="" gt="REG"/>
      <g eid="2018122313" gsis="57804" d="Sun" t="4:25" q="F" k="" h="NO" hnn="saints" hs="31" v="PIT" vnn="steelers" vs="28" p="" rz="" ga="" gt="REG"/>
      <g eid="2018122314" gsis="57805" d="Sun" t="8:20" q="F" k="" h="SEA" hnn="seahawks" hs="38" v="KC" vnn="chiefs" vs="31" p="" rz="" ga="" gt="REG"/>
      <g eid="2018122400" gsis="57806" d="Mon" t="8:15" q="F" k="" h="OAK" hnn="raiders" hs="27" v="DEN" vnn="broncos" vs="14" p="" rz="" ga=""
      gt="REG"/>
    </gms>
</ss>
```

Weekly Games Reference

Query:

- http://www.nfl.com/ajax/scorestrip?season=**2018**&seasonType=**POST**&week=**20**


```
-<ss>
 -<gms gd="0" w="20" y="2018" t="P">
    <g eid="2019012000" gsis="57831" d="Sun" t="3:05" q="FO" k="" h="NO" hnn="saints" hs="23" v="LA" vnn="rams" vs="26" p="" rz="" ga="" gt="CON"/>
    <g eid="2019012001" gsis="57832" d="Sun" t="6:40" q="FO" k="" h="KC" hnn="chiefs" hs="31" v="NE" vnn="patriots" vs="37" p="" rz="" ga="" gt="CON"/>
  </gms>
 </ss>
```

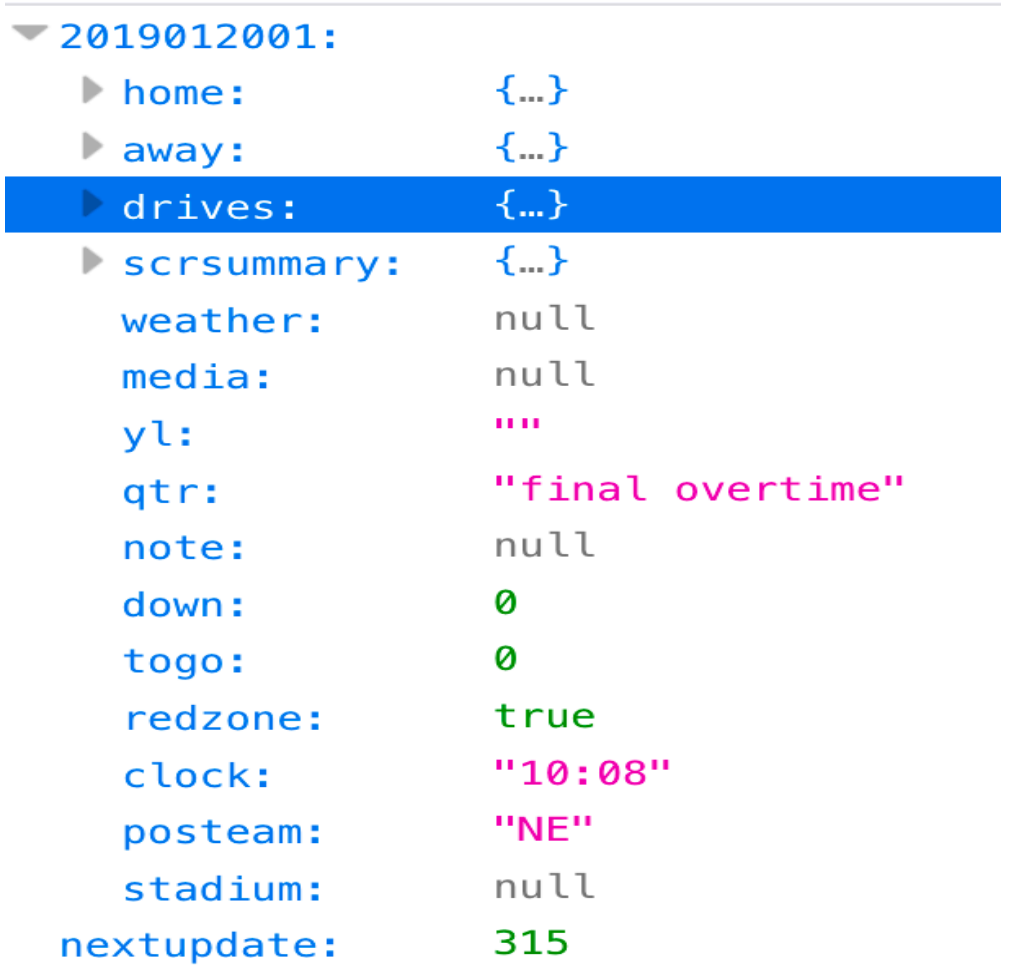
Game Data Raw

{"2019012001":{"home":{"score":{"1":0,"2":0,"3":7,"4":24,"5":0,"T":31},"abbr":"KC","to":3,"stats":{"passing":{"00-0033873":{"name":"P.Mahomes","att":31,"cmp":16,"yds":295,"tds":3,"ints":0,"twopta":0,"twoptm":0}},"rushing":{"00-0030874":{"name":"Dam.
Williams","att":10,"yds":30,"tds":1,"lng":10,"lngtd":2,"twopta":0,"twoptm":0},***"00-0033873":{"name":"P.Mahomes","att":2,"yds":11,"tds":0,"lng":9,"lngtd":0,"twopta":0,"twoptm":0}},"receiving***":{"00-0030874":{"name":"Dam.
Williams","rec":5,"yds":66,"tds":2,"lng":33,"lngtd":23,"twopta":0,"twoptm":0},"00-0031325":{"name":"S.Watkins","rec":4,"yds":114,"tds":0,"lng":54,"lngtd":0,"twopta":0,"twoptm":0},"00-
0030506":{"name":"T.Kelce","rec":3,"yds":23,"tds":1,"lng":12,"lngtd":12,"twopta":0,"twoptm":0},"00-0033040":{"name":"T.Hill","rec":1,"yds":42,"tds":0,"lng":42,"lngtd":0,"twopta":0,"twoptm":0},"00-
0032775":{"name":"D.Robinson","rec":1,"yds":27,"tds":0,"lng":27,"lngtd":0,"twopta":0,"twoptm":0},"00-0030414":{"name":"S.Ware","rec":1,"yds":21,"tds":0,"lng":21,"lngtd":0,"twopta":0,"twoptm":0},"00-
0030155":{"name":"D.Harris","rec":1,"yds":2,"tds":0,"lng":2,"lngtd":0,"twopta":0,"twoptm":0}},"fumbles":{"00-0033873":{"name":"P.Mahomes","tot":1,"rcv":1,"trcv":1,"yds":0,"lost":0}},"kicking":{"00-
0033303":{"name":"H.Butker","fgm":1,"fga":1,"fgyds":39,"totpfg":3,"xpmade":4,"xpmissed":0,"xpa":4,"xpb":0,"xptot":4}},"punting":{"00-0023534":{"name":"D.Colquitt","pts":5,"yds":217,"avg":36,"i20":1,"lng":59}},"kickret":{"00-0034278":{"name":"Tr.
Smith","ret":4,"avg":24,"tds":0,"lng":29,"lngtd":0},"00-0030155":{"name":"D.Harris","ret":1,"avg":17,"tds":0,"lng":17,"lngtd":0}},"puntret":{"00-0033040":{"name":"T.Hill","ret":1,"avg":-11,"tds":0,"lng":-11,"lngtd":0}},"defense":{"00-
0030665":{"name":"D.Sorensen","tkl":11,"ast":3,"sk":0.0,"int":1,"ffum":0},"00-0031348":{"name":"A.Hitchens","tkl":7,"ast":7,"sk":0.0,"int":0,"ffum":0},"00-0033058":{"name":"R.Ragland","tkl":7,"ast":5,"sk":0.0,"int":1,"ffum":0},"00-
0033084":{"name":"K.Fuller","tkl":7,"ast":3,"sk":0.0,"int":0,"ffum":0},"00-0034573":{"name":"C.Ward","tkl":4,"ast":3,"sk":0.0,"int":0,"ffum":0},"00-0031713":{"name":"X.Williams","tkl":3,"ast":4,"sk":0.0,"int":0,"ffum":0},"00-
0027858":{"name":"E.Berry","tkl":5,"ast":1,"sk":0.0,"int":0,"ffum":0},"00-0034818":{"name":"D.Nnadi","tkl":0,"ast":6,"sk":0.0,"int":0,"ffum":0},"00-0028024":{"name":"A.Bailey","tkl":1,"ast":4,"sk":0.0,"int":0,"ffum":0},"00-
0032150":{"name":"S.Nelson","tkl":3,"ast":0,"sk":0.0,"int":0,"ffum":0},"00-0033038":{"name":"E.Murray","tkl":2,"ast":0,"sk":0.0,"int":0,"ffum":0},"00-0028008":{"name":"J.Houston","tkl":2,"ast":0,"sk":0.0,"int":0,"ffum":0},"00-
0032797":{"name":"J.Lucas","tkl":0,"ast":2,"sk":0.0,"int":0,"ffum":0},"00-0034774":{"name":"B.Speaks","tkl":0,"ast":2,"sk":0.0,"int":0,"ffum":0},"00-
0031636":{"name":"J.Hamilton","tkl":0,"ast":1,"sk":0.0,"int":0,"ffum":0}},"team":{"totfd":18,"totyds":290,"pyds":249,"ryds":41,"pen":4,"penyds":28,"trnovr":0,"pt":5,"ptyds":217,"ptavg":36,"top":"20:53"}},"players":null},"away":{"score":{"1":7,"2":7,"3":3,"4":14,"
5":6,"T":37},"abbr":"NE","to":3,"stats":{"passing":{"00-0019596":{"name":"T.Brady","att":46,"cmp":30,"yds":348,"tds":1,"ints":2,"twopta":0,"twoptm":0}},"rushing":{"00-
0034845":{"name":"S.Michel","att":29,"yds":113,"tds":2,"lng":11,"lngtd":10,"twopta":0,"twoptm":0},"00-0030288":{"name":"R.Burkhead","att":12,"yds":41,"tds":2,"lng":14,"lngtd":4,"twopta":0,"twoptm":0},"00-
0031062":{"name":"J.White","att":6,"yds":23,"tds":0,"lng":9,"lngtd":0,"twopta":0,"twoptm":0},**"00-0019596":{"name":"T.Brady","att":1,"yds":-1,"tds":0,"lng":-1,"lngtd":0,"twopta":0,"twoptm":0}},"receiving**":{"00-
0027150":{"name":"J.Edelman","rec":7,"yds":96,"tds":0,"lng":20,"lngtd":0,"twopta":0,"twoptm":0},**"00-0027656":{"name":"R.Gronkowski","rec":6,"yds":79,"tds":0,"lng":25,"lngtd":0,"twopta":0,"twoptm":0}**,"00-
0028237":{"name":"C.Hogan","rec":5,"yds":45,"tds":0,"lng":11,"lngtd":0,"twopta":0,"twoptm":0},"00-0031062":{"name":"J.White","rec":4,"yds":49,"tds":0,"lng":30,"lngtd":0,"twopta":0,"twoptm":0},"00-
0030288":{"name":"R.Burkhead","rec":4,"yds":23,"tds":0,"lng":6,"lngtd":0,"twopta":0,"twoptm":0},"00-0030578":{"name":"C.Patterson","rec":2,"yds":18,"tds":0,"lng":15,"lngtd":0,"twopta":0,"twoptm":0},"00-
0032208":{"name":"P.Dorsett","rec":1,"yds":29,"tds":1,"lng":29,"lngtd":29,"twopta":0,"twoptm":0},"00-0027925":{"name":"J.Develin","rec":1,"yds":9,"tds":0,"lng":9,"lngtd":0,"twopta":0,"twoptm":0}},"fumbles":{},"kicking":{"00-
0024333":{"name":"S.Gostkowski","fgm":1,"fga":1,"fgyds":47,"totpfg":3,"xpmade":4,"xpmissed":0,"xpa":4,"xpb":0,"xptot":4}},"punting":{"00-0029984":{"name":"R.Allen","pts":2,"yds":75,"avg":43,"i20":2,"lng":43}},"kickret":{"00-
0030578":{"name":"C.Patterson","ret":3,"avg":26,"tds":0,"lng":38,"lngtd":0},"00-0027925":{"name":"J.Develin","ret":1,"avg":2,"tds":0,"lng":2,"lngtd":0}}, . . .

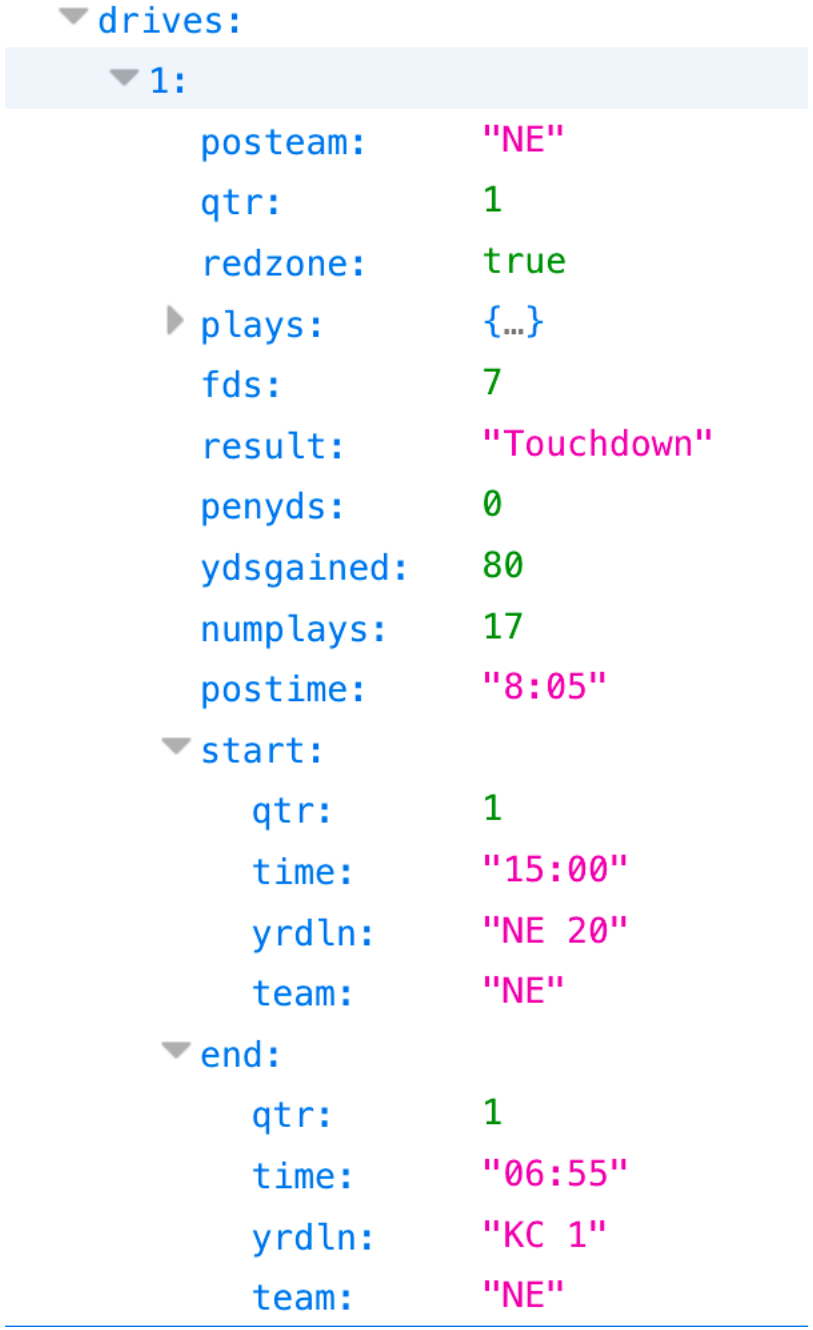
Game Data JSON

Query:

- http://www.nfl.com/liveupdate/game-center/**2019012001**/**2019012001**_gtd.json



Game Data JSON



Game Data JSON

Game Data JSON


```
105:
    sp:       0
    qtr:      1
    down:     2
    time:     "13:40"
    yrdln:    "NE 32"
    ydstogo:  9
    ydsnet:   80
    posteam:  "NE"
    desc:     "(13:40) T.Brady pass short middle to R.Gronkowski to NE 45 for 13 yards (D.Sorensen)."
    note:     null
    players:  {…}
```

Homework Week 1

- Watch Super Bowl LIII
- Record all drives of one full quarter play by play